ELECTRA와 Label Attention Network를 이용한 한국어 개체명 인식

김홍진[O,1], 오신혁[1], 김학수[2]

강원대학교 컴퓨터정보통신공학전공[1], 건국대학교 인공지능학과[2]
jin3430@kangwon.ac.kr, osh7605@kangwon.ac.kr, nlpdrkim@konkuk.ac.kr

Korean Named Entity Recognition Using ELECTRA and Label Attention Network

Hong-Jin Kim[O,1], Shin-Hyeok Oh[1], Hark-Soo Kim[2]
Kangwon National University Department of Computer and Communications Engineering[1],
Konkuk University Department of Artificial Inteligence[2]

요 약

개체명 인식이란 문장에서 인명, 지명, 기관명 등과 같이 고유한 의미를 갖는 단어를 찾아 개체명을 분류하는 작업이다. 딥러닝을 활용한 연구가 수행되면서 개체명 인식에 RNN(Recurrent Neural Network)과 CRF(Condition Random Fields)를 결합한 연구가 좋은 성능을 보이고 있다. 그러나 CRF는 시간 복잡도가 분류해야 하는 클래스(Class) 개수의 제곱에 비례하고, 최근 RNN과 Softmax 모델보다 낮은 성능을 보이는 연구도 있었다. 본 논문에서는 CRF의 단점을 보완한 LAN(Label Attention Network)와 사전 학습 언어 모델인 음절 단위 ELECTRA를 활용하는 개체명 인식 모델을 제안한다.

주제어: 개체명 인식, 음절 단위, 언어 모델, Label Attention Network

1. 서론

개체명 인식(Named Entity Recognition)은 입력 문장에서 인명, 지명, 기관명 등과 같이 고유한 의미를 갖는 단어를 찾아 개체명을 분류하는 작업이다. 대부분의 개체명 인식 연구에서는 개체명 인식 문제를 순차적 레이블링(Sequence Labeling) 문제로 간주하여 해결했다. 최근 딥러닝을 활용한 연구가 활발히 수행되면서 순차적 레이블링 문제에 RNN(Recurrent Neural Network) 기반 계층과 CRF(Condition Random Fields)를 결합한 연구가 좋은 성능을 보였다. 그러나 CRF의 시간 복잡도는 분류해야 하는 클래스(Class) 개수의 제곱에 비례하기 때문에 세부 분류 개체명과 같이 태그 개수가 많은 작업에는 적합하지 않다. 또한 순차적 레이블링으로 해결 가능한 형태소 분석에서 단순히 RNN에 Softmax만을 적용한 모델이 RNN과 CRF를 결합한 모델 보다 좋은 성능을 보이는 연구도 있었다[1]. 본 논문에서는 CRF의 단점을 보완한 LAN(Label Attention Network)를 활용하여 음절 단위 한국어 개체명 인식 모델을 제안한다.

2. 관련 연구

최근 개체명 인식 연구에는 딥러닝을 활용하는 방법이 주를 이루고 있다. 특히 딥러닝 기법 중 순차적 레이블링에 좋은 성능을 보이는 양방향(Bidirectional) RNN과 CRF를 결합한 모델인 BiLSTM(Long Short-Term Memory)-CRF 또는 BiGRU(Gated Recurrent Unit)-CRF 모델이 개체명 인식에서 좋은 성능을 보였다[2-5]. 한편, 대용량 말뭉치를 이용해 학습한 언어 모델이 다양한 자연어 처리 분야에서 좋은 성능을 보이고 있다. 한국어 개체명 인식에서도 언어 모델을 대용량 말뭉치로 학습시킨 후 활용한 연구가 진행되었다[6-8]. [6]은 트랜스포머(Trnasformer)[9]와 셀프 어텐션 매커니즘(Self-Attention Mechanism)을 이용하여 문장에서 임의로 단어를 마스킹(Masking)하고 예측하도록 학습한 BERT(Bidirectional Encoder Representation from Transformers)[10]를 활용하여 개체명 인식을 수행했다. [7]은 BERT와 달리 임의의 단어가 매 학습마다 동적으로 마스킹 되도록 개선한 RoBERTa[11]를 활용하여 개체명 인식을 수행했다. [8]은 단어 임베딩(Embedding) 파라미터(Parameter)의 크기 부담을 줄이고, 기존 BERT 모델의 파라미터 공유를 통해 학습 효율을 높인 ALBERT[12]를 활용하여 개체명 인식을 수행했다. BERT의 후속 연구로 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)[13]는 Generator에서 임의의 단어를 마스킹하고 예측하도록 학습한 다음, Discriminator에서 생성한 단어 열에 대해서 각 단어가 원래 입력과 동일한 것인지 치환된 것인지 예측하도록 학습한다. 본 논문에서는 음절 표현을 위해 음절 단위 한국어 ELECTRA를 이용하고 LAN을 결합하여 음절 단위 개체명 인식을 수행한다.

3. ELECTRA와 LAN을 이용한 개체명 인식 모델

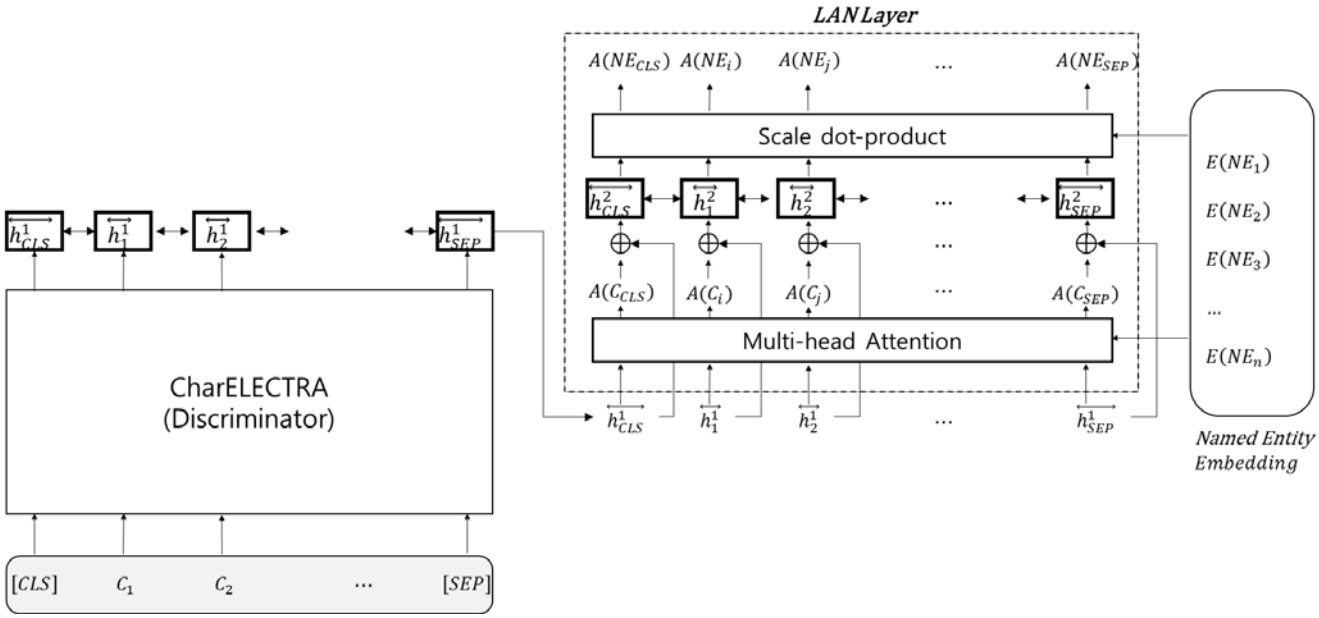

그림 1 개체명 인식 모델 전체 구조도

그림 1은 본 논문에서 제안하는 개체명 인식 모델의 전체 구조이며 ELECTRA, LAN 계층으로 구성되어 있다. ELECTRA에서 입력 데이터를 받아 음절 표현 벡터(Vector)를 생성하고 LAN 계층에서 음절 표현 벡터를 이용하여 각 음절과 개체명 임베딩과의 Multi-head Attention Mechanism[9]을 통해 가장 연관성 있는 개체명을 출력한다.

3.1 LAN 계층

LAN 계층에서는 ELECTRA의 출력인 각 음절 벡터와 개체명 레이블 임베딩 간의 연관성을 계산하여 각 음절에 대한 개체명을 예측한다. 그림 1에서 C_i는 ELECTRA의 출력인 음절 벡터이다. 첫번째 계층에서는 문맥 정보를 반영하기 위해 음절 벡터를 양방향 LSTM으로 인코딩한 후 연결하여 사용하며 수식은 다음과 같다.

$$\begin{gathered}\overrightarrow{h_i^1} = LSTM(C_i, \overrightarrow{h_{i-1}^1}) \\ \overleftarrow{h_i^1} = LSTM(C_i, \overleftarrow{h_{i-1}^1}) \\ \overleftrightarrow{h_i^1} = [\overrightarrow{h_i^1}; \overleftarrow{h_i^1}] \\ \overrightarrow{H^1} = \{\overleftrightarrow{h_{CLS}^1}, \overleftrightarrow{h_1^1}, \dots, \overleftrightarrow{h_{SEP}^1}\}\end{gathered} \quad (1)$$

수식 (1)에서 $\overrightarrow{h_i^1}$는 정방향 은닉 상태(Forward Hidden State)이며 $\overleftarrow{h_i^1}$는 역방향(Backward) 은닉 상태이다. $\overleftrightarrow{h_i^1}$는 i번째 단계에서 양방향 은닉 상태를 연결한 벡터이다. $\overrightarrow{H^1}$는 양방향 문맥 정보가 반영된 각 단어를 나타내는 벡터이다. 다음으로, $\overrightarrow{H^1}$과 그림 1의 개체명 임베딩인 $E(NE) = \{E(NE_1), E(NE_2), \dots, E(NE_n)\}$ 사이의 연관성을 계산하기 위해 Multi-head Attention을 사용하며 수식은 다음과 같다.

$$\begin{gathered}head_j = Attention(QW_j^Q, KW_j^K, VW_j^V) = \alpha_j * VW_j^V \\ Q = \overrightarrow{H^1}, K = V = E(NE) \\ \alpha_j = softmax(\frac{QW_j^Q * (KW_j^K)^T}{\sqrt{d_h}}) \\ A(C_i) = [head_1; head_2; \dots; head_k]\end{gathered} \quad (2)$$

수식 (2)에서 W_j^Q, W_j^K, W_j^V은 k개의 head 중 j번째 head의 가중치 파라미터이며, 사용한 개체명 임베딩은 랜덤 초기화(Random Initialize)하여 사용한다. 가중치 파라미터와 개체명 임베딩은 학습 과정에서 미세 조정(Fine-tuned)된다. d_h는 정규화 값이며 개체명 임베딩 크기와 동일하다. 또한 $A(C_i)$는 어텐션 벡터이며 각 음절에 대한 개체명 분포가 강조된 벡터를 나타낸다. 두번째 계층에서는 첫번째 LSTM의 출력인 $\overrightarrow{H^1}$와 어텐션 벡터 $A(C_i)$를 연결(Concatenation)하여 양방향 LSTM으로 인코딩하며 수식은 다음과 같다.

$$\begin{gathered}\overrightarrow{h_i^2} = LSTM([\overrightarrow{H^1}; A(C_i)], \overrightarrow{h_{i-1}^2}) \\ \overleftarrow{h_i^2} = LSTM([\overrightarrow{H^1}; A(C_i)], \overleftarrow{h_{i-1}^2}) \\ \overleftrightarrow{h_i^2} = [\overrightarrow{h_i^2}; \overleftarrow{h_i^2}] \\ \overrightarrow{H^2} = \{\overleftrightarrow{h_{CLS}^2}, \overleftrightarrow{h_1^2}, \dots, \overleftrightarrow{h_{SEP}^2}\}\end{gathered} \quad (3)$$

다음으로, $\overrightarrow{H^2}$과 $E(NE)$에 대하여 Scaled dot-Product[9]를 수행하며 수식은 다음과 같다.

$$\begin{gathered}A(NE_i) = \frac{Q * K^T}{\sqrt{d_h}} \\ Q = \overrightarrow{H^2}, K = E(NE)\end{gathered} \quad (4)$$

수식 (4)에서 계산된 각 음절 벡터와 개체명 임베딩의

내적 값인 $A(NE_i)$를 이용하여 개체명은 다음과 같이 예측된다.

$$\widehat{NE} = argmax(softmax(A(NE_i))) \quad (5)$$

3.2 학습 방법

본 논문에서는 학습을 위해 모델이 예측한 각 음절의 개체명과 정답 개체명 간의 크로스 엔트로피를 최소화하도록 학습하며 수식은 다음과 같다.

$$H_{\widehat{NE}} = -\sum_i \widehat{NE}_i \log(NE_i) \quad (6)$$

4. 실험 및 결과

본 논문에서 사용한 음절 단위 ELECTRA는 20GB의 한국어 위키피디아, 뉴스 데이터를 사용하여 사전 학습한 모델이다. 또한 실험 데이터로 ETRI의 엑소브레인 언어분석 말뭉치를 이용하였다. 학습 데이터는 9만개, 평가 데이터는 1만개로 구성되어 있다. 개체명 종류는 총 15개로 다음 표 1과 같다.

표 1 개체명 종류

개체명 분류	표기
인물	PS
지역	LC
기관	OG
인공물	AF
날짜	DT
시간	TI
문명	CV
동물	AM
식물	PT
수량	QT
학문분야	FD
이론	TR
사건	EV
물질	MT
용어	TM

표 2은 본 논문에서 사전 학습한 음절 단위 ELECTRA와 결합하는 각 모델에 따른 ETRI 데이터의 성능 비교이다.

표 2 모델에 따른 성능 비교

Model	F1-score
ELECTRA + LSTM + Softmax	0.9249
ELECTRA + LSTM + CRF	0.9251
ELECTRA + LAN	**0.9278**

표 2에서 ELECTRA+ LSTM+ Softmax는 ELECTRA의 출력인 음절 벡터를 양방향 LSTM으로 인코딩한 후 Softmax를 이용하여 개체명 인식을 수행한 모델이며, ELECTRA+ LSTM+ CRF은 ELECTRA 출력을 양방향 LSTM으로 인코딩 후 CRF를 이용하여 개체명 인식을 수행한 모델이다. ELECTRA+ LAN은 본 논문에서 제안하는 방법인 ELECTRA의 출력을 양방향 LSTM으로 인코딩 후 Multi-head Attention 기법으로 개체명 임베딩과 각 음절의 연관성을 계산하여 개체명 인식을 수행한 모델이다.

실험 결과, 본 논문에서 제안한 ELECTRA와 LAN을 결합한 모델이 가장 좋은 성능을 보였다.

표 3은 제안 모델과 동일한 데이터를 사용한 기존 모델의 성능 비교이다.

표 3 기존 모델과 제안 모델의 성능 비교

Model	F1-score
BERT + CRF[6]	0.9158
ALBERT + LSTM + CRF[8]	0.9187
RoBERTa + LSTM + CRF[7]	0.9194
ELECTRA + LAN(Ours)	**0.9278**

표 3에서 [6]은 BERT의 출력 값에 CRF를 적용한 모델이다. [7, 8]은 각각 RoBERTa와 ALBERT의 출력 값을 양방향 LSTM으로 인코딩 한 후 CRF를 적용한 모델이다. 실험 결과, 본 논문에서 제안한 ELECTRA와 LAN을 결합한 모델이 가장 좋은 성능을 보였다.

표 4는 LAN과 CRF의 문장 처리 속도 비교이다.

표 4 LAN과 CRF의 속도 비교

Model	문장 처리 개수(문장/초)
ELECTRA + LSTM + CRF	195
ELECTRA + LAN	**260**

표 4에서 CRF 모델은 초당 195개의 문장을 처리할 수 있고, LAN 모델은 초당 260개의 문장을 처리할 수 있다. 따라서 LAN 모델이 CRF 모델보다 효율적인 시간 복잡도를 보인다.

5. 결론 및 향후연구

본 논문에서는 언어 모델인 ELECTRA를 음절 단위의 한국어 대용량 말뭉치로 사전 학습하여 활용하고, LAN을 결합하여 기존에 개체명 인식에서 가장 높은 성능을 보이던 LSTM과 CRF를 결합한 모델 보다 좋은 성능을 보인다. 향후 연구로 세부 분류 개체명 인식을 수행하여 LAN과 CRF에 대하여 개체명 종류가 증가함에 따른 시간 분석도 및 성능을 비교하는 실험을 진행할 예정이다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국

연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1069737).

참고문헌

[1] C. Leyang and Y. Zhang, "Hierarchically-Refined Label Attention Network for Sequence Labeling", Empirical Methods in Natural Language Processing (EMNLP) and International Joint Conference on Natural Language Processing (IJCNLP), pp. 4106-4119, 2019.

[2] 나승훈, 민진우, "문자 기반 LSTM CRF를 이용한 개체명 인식", 한국정보과학회 2016년 한국컴퓨터종합학술대회 논문집, pp. 729-731, 2016.

[3] 유홍연, 고영중, "Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장", 정보과학회 논문지, 제44권, 제3호, pp. 306-313, 2017.

[4] 박건우, 박성식, 장영진, 최기현, 김학수, "KACTEIL-NER: 딥러닝과 앙상블 기법을 이용한 개체명 인식기", 제29회 한글 및 한국어 정보처리 학술대회 논문집, pp. 324-326, 2017.

[5] 김홍진, 김학수 "딥러닝 기반의 개체명 인식을 위한 효과적인 사전 자질 사용 방법", 제31회 한글 및 한국어 정보처리 학술대회 논문집, pp. 293-296, 2019.

[6] 박관형, 나승훈, 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정", 한국정보과학회 한국소프트웨어종합학술대회 논문집, pp. 584-586, 2019.

[7] 민진우, 나승훈, 신종훈, 김영길, "RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱", 한국정보과학회 한국소프트웨어종합학술대회 논문집, pp. 407-409. 2019.

[8] 이영훈, 나승훈, 최윤수, 이혜우, 장두성, "ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해", 한국정보과학회 한국소프트웨어종합학술대회 논문집, pp. 332-334. 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all You Need", Neural Information Processing Systems (NIPS), pp. 5998-6008, 2017.

[10] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (1), 2019.

[11] Y. Liu, M. Ott, Na. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arxiv.org/abs/1907.11692, 2019.

[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations", in: Proceedings of ICLR, 2020.

[13] K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", International Conference on Learning Representations (ICLR), 2019.